Searching corpora of Chinese and British writers for lexicalised language

Searching corpora of Chinese and British writers for lexicalised language
中英学生写作语料库中惯用语使用的探析

Maria Leedham

The Open University

Outline

- Study overview and RQs
- Lexicalised language
- Comparing the writing
- Questionnaire data

The Study

My Background:

- Research Assistant on a corpus project collecting student assignments and interviewing lecturers, ('British Academic Written English' project)
- English for Academic Purposes teacher in universities teaching Chinese students.

- My PhD looks at Chinese and British students' assessed writing in U.K. universities.

- Years 1,2,3 & PG; narrow to 3 discipline areas; corpus study plus questionnaire and interviews.

Research Questions

- 1. In what ways do British and Chinese undergraduate and Master's students in U.K. universities differ in their use of lexicalised language in the academic writing of three selected disciplines?

- 2. How do these students develop their use of lexicalised language from year 1 to year 3 of undergraduate study?

- 3. What are the pedagogical implications for teachers of academic writing?

Building the Corpus 1

- A corpus is a balanced collection of texts, usually stored in electronic form.
- My two corpora will comprise 300 assignments from Chinese students and 300 from British students – around 1 million words per corpus.
- Assignments will be matched for discipline, year groups, age and gender of students. All assignments will be anonymised.
- The two corpora can be searched using corpus linguistics software e.g. WordSmith Tools v.5

Building the Corpus 2

- British Academic Written English corpus (Bawe) as a starting point :-

  - 4 Universities (Warwick, Reading, Oxford Brookes & Coventry)
  - 32 different disciplines
  - 3000 assignments
  - II:i or I level.

- Currently – collecting more assignments (70 so far) through contacts, CSSA, FaceBook.

Overview of Chinese corpus

	Year 1	Year 2	Year 3	Master's	Total per discipline
Life Sciences (food, biological & plant sciences)	21	17	15	17	**70**
Social Sciences (business, economics and hospitality)	19	11	52	25	**107**
Physical Sciences (cybernetics, all engineering, computing)	12	16	48	5	**81**
Arts & Humanities (linguistics, ICT in education, theatre studies)	0	0	5	16	**21**
Totals per year group	**52**	**44**	**120**	**63**	**Total: 279**

Outline

- Study overview and RQs
- Lexicalised language
- Comparing the writing
- Questionnaire data

Lexicalised language as ‘chunks’

- Psychology research suggests humans group phenomena together, i.e. we process information in "chunks" or meaningful units of information (Miller, 1956).

  e.g. chess (remembering sequences of moves as chunks), telephone numbers, stretches of music

- Thus - “it is possible to expand the total amount of information by packing more and more information into one chunk". (Howard 1983:104).

  e.g. A chess expert has sequences of longer moves than the novice player; a skilled language user chooses from a variety of long lexicalized sentence stems while a lower-level user struggles to communicate with a few short chunks.

- So - we use “an abundant resource (memory to store prefabricated chunks of language) to compensate for a limited one (processing capacity)”.

  (Schmitt and McCarthy, 1997:230).

Examples of lexicalised language or ‘chunks’

- it is interesting to note
- in other words
- in order to
- and so on
- on the other hand
- the fact is
- it has been argued that
- based on the
- human beings are
- is there a link between
- such as
- individual needs
- are there
- out of

How much language is lexicalised?

4-5% Moon , 1998 "FEI"= fixed expression and idiom

"holistic units of two or more words"

Includes frozen collocations, proverbs, routine formulae, sayings, similes

e.g. *armed to the teeth, foot the bill, red herring, at home, little by little,*

Excludes compound nouns, adjectives, verbs; phrasal verbs, foreign phrases, multi-word inflectional forms e.g. *civil servant, self-raising, had been lying, more careful*

80% Altenberg, 2001 "recurrent word combinations"

Includes "any continuous string of words occurring more than once in identical form"

e.g. *I think that, do you know, out of the, what sort of, because I mean, and then I,*

Excludes any word strings occurring once only.

How do we find lexicalised language?

- ScriptLog software
- Eye-tracking

ScriptLog with Eyetracker

- Structural features

  e.g. fixedness, non-compositionality

- Phonological features,

  e.g. speech rate, clarity of articulation, intonation contour, lack of internal pausing

- Frequency counts
- Human intuition

Lexicalised Language

From frequency counts

- 'Clusters' are "words which are found repeatedly together in each others' company, in sequence".

  Mike Scott, 2004, WordSmith Tools.

- Frequently occur across structural groups

  e.g. *way we speak is there a*

- Evidence to suggest that not all of these clusters are stored as wholes in the mental lexicon.

From human intuition

- A 'formulaic sequence' is "a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar"

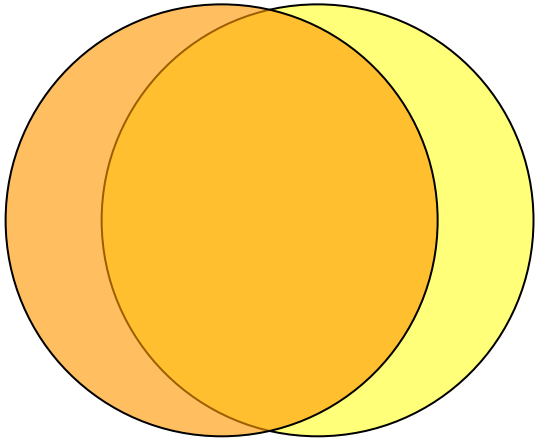
  (Alison Wray, 2002:9). Also Schmitt, 2004.

- Sequences do not cross structural boundaries - e.g. *the way we speak*
- Sequences are "psychologically real" and are thought to be stored as wholes in the mental lexicon.

Lexicalised Language

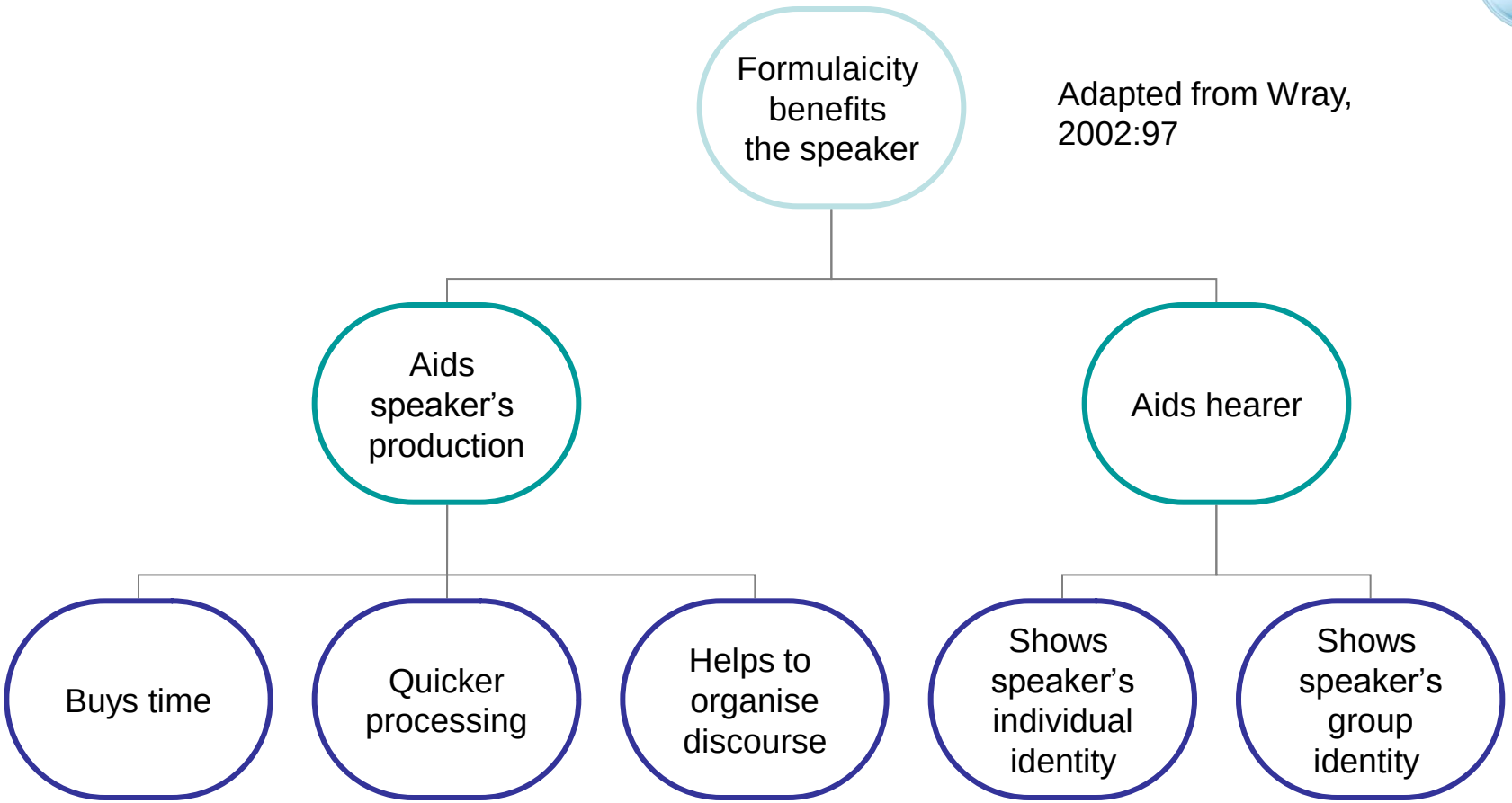
Lexicalisation = “the process by which a string of words and morphemes becomes institutionalized as part of the language and develops its own specialist meaning or function.” Moon, 1998:36.

Clusters



Formulaic Sequences

Functions of lexicalised language



Adapted from Wray, 2002:97

"The more novel our output is for the hearer, the more likely it is to be misunderstood" (Wray, 2002:94).

Outline

- Study overview and RQs
- Lexicalised language
- Comparing the writing
- Questionnaire data

ChiCor

No.	Chunk	Freq.	Texts
1	on the other hand	61	42
2	as a result of	26	19
3	as well as the	24	17
4	at the same time	22	15
5	one of the most	20	13
6	it is important to	19	17
7	as one of the	15	13
8	at the end of	14	9
9	it is necessary to	14	11
10	can be used to	13	8
11	it is difficult to	13	10
12	can be seen that	12	10
13	it can be seen	12	9
14	it is believed that	12	10
15	this is due to	11	9

EngCor

No.	Chunk	Freq.	Texts
1	at the end of	28	21
2	as a result of	25	17
3	it is important to	19	17
4	at the same time	14	13
5	can be used to	14	13
6	it can be seen	14	9
7	this is due to	14	11
8	can be seen that	13	8
9	it is possible to	11	8
10	on the other hand	11	8
11	it is clear that	10	10
12	may be due to	10	8

It is vital for motivation theories to consider four influential individual needs in the work place, which are the competence, achievement, affiliation, and money motives. 'The competence motive' is the desire for job mastery and professional growth. Robert White suggests the competence motive to be based on the assumption that a person is not only "a vehicle for a set of instincts" (Gellerman 1963: 111), but is also eager on discovering and fulfilling their potential. It is assumed that humans are keen on manipulating their environment to pursue goals. Thus, competence is a key motive affecting job success, because people who have faith in their own ability to influence the environment do tend to succeed.

On the other hand, individuals with a strong achievement motive perceive accomplishment as an ends. Achievement-motivated employees search for the opportunities to obtain successes that are "hard but are not unobtainable" (Gullerman 1965: 126), and thus, tend to outperform others by constantly challenging themselves. The reasonable degree of risk involved in the goal-attainment process encourages employees to set realistic goals and to maximize their abilities.

Affiliation is another individual need, which refers to the "social drive to be associated with others in interdependent relationships, involving using others for help or support without making them responsible for problems" (MerckSource, 2006). Affiliation can be considered as a means to an end or an ends itself – people socialize with fellow workers for specific purposes, such as favors or protection, or simply for enjoyment (Gellerman, 1965).

Concordance lines for ‘on the other hand’

ChiCor

No.	Concordance	Search term	
1	increased fat deposition with age.	On the other hand,	males tend to have
2	and reduces the cost dramatically.	On the other hand,	Ford’s product still
3	y change when the product is used.	On the other hand,	it’s really difficult
4	cell’s viewing angle performance.	On the other hand,	the viewing

EngCor

1	by (Williams and Feltmate, 1992).	On the other hand,	if there is a form of
2	ates smoking tobacco and cancer.	On the other hand,	the drug testing for
3	n in Fig 14 and 15) formica fusca,	on the other hand,	was found mainly in
4	very much an altruistic view. Then	on the other hand,	is it ethical to stop

N	Word	Freq.	Text	%
1	it can be seen that the	15	12	3.39
2	it is important to note that	15	8	2.26
3	set of tasks questions exercises non	13	13	3.67
4	gender on the way we speak	12	6	1.69
5	is there a link between them	12	6	1.69
6	speak is there a link between	12	6	1.69
7	the social variables of class and	12	5	1.41
8	the way we speak is there	12	6	1.69
9	way we speak is there a	12	6	1.69
10	we speak is there a link	12	6	1.69
11	and gender on the way we	11	6	1.69
12	class and gender on the way	11	6	1.69
13	dot plot to show the mean	10	2	0.56
14	is due to the fact that	10	10	2.82
15	of class and gender on the	10	5	1.41
16	of the social variables of class	10	5	1.41
17	signs but do they have language	10	6	1.69
18	variables of class and gender on	10	5	1.41
19	in order to be able to	9	7	1.98

Concordance lines for “the way we speak”

- 6018b What is the effect of the social variables of gender and class on the way we speak? Is there a link between them?
- 6018b The effects of gender and class on the way we speak is a question that has engaged much time with linguists and
- 6018b As well as social class, Gender also obviously has a huge effect on the way we speak - both in single sex and mixed sex
- 6018b way we speak as well as helping us link the effects of social class and gender on the way we speak. Lakoff has claimed

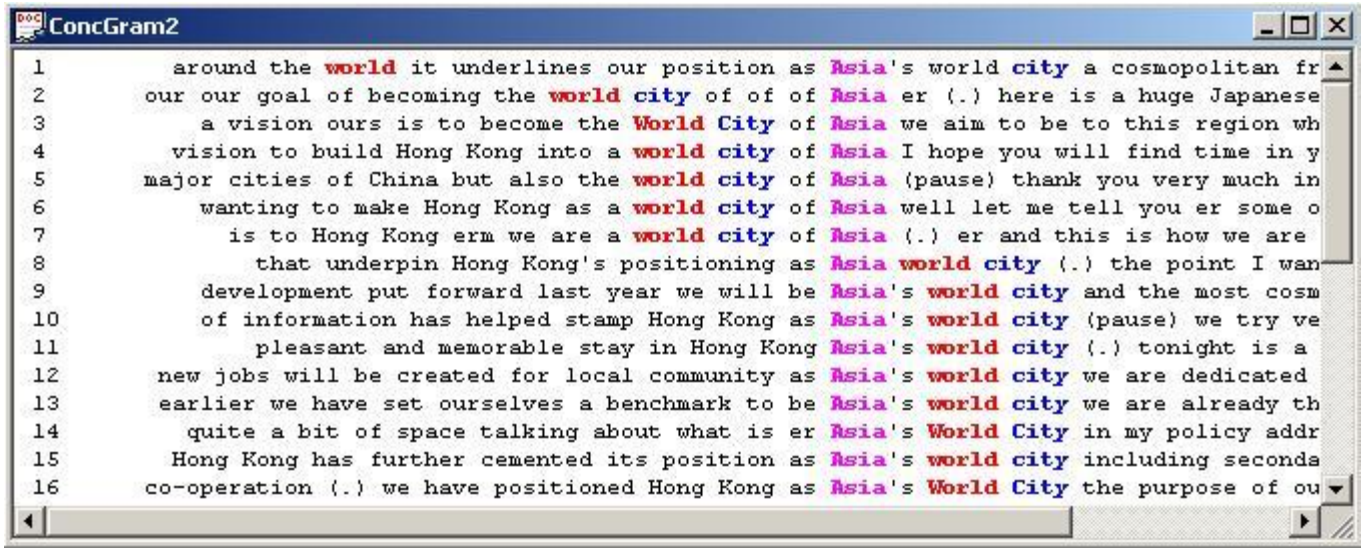
- 6055b What is the effect of the social variables of class and gender on the way we speak? Is there a link between them?
- 6055b I will initially discuss the effect of social class on the way we speak and followed by gender as the other social variable.

- *<teiHeader><fileDesc><titleStmt><title>*

  What is the effect of the social variables of gender and class on the way we speak? Is there a link between them?

Beyond clusters?

- Clusters, n-grams, = contiguous words
- ConcGrams, Greaves & Warren, 2007, e.g.A**B, B*A,



PoS-gram
e.g. prep.+det.+noun+of
at the end of, as a result of,

Semantic sequences
e.g. time + journey time + transport
In Summer it's a half-hour journey by bus

Outline

- Study overview and RQs
- Lexicalised language
- Comparing the writing
- Questionnaire data

Questionnaire Data

- 170 questionnaire responses so far to online survey of British and Chinese students' view on writing assignments.
- Over 40 universities and over 20 disciplines are represented.

- Where and how were you taught academic writing?
- At secondary school...We were taught vocabulary to use. (year 2, *Cantonese speaker)*
- I have never been taught about essay writing ...

  I still don't even know if I am doing it right but as I have passed so far I guess it's ok. *(year 3, native English speaker)*

Have you changed your way of planning and writing your assignments?

- I had been used to writing in Chinese first and then translating. But afterward I switched to writing and planning at the same time... *(Mandarin)*
- Yes - I plan them before I write. Also, I've absorbed some of the 'style' appropriate to the discipline. *(NES)*
- I plan, write a draft, leave it, then go back and re-read and adjust it. repeat as necessary. *(NES)*

Generally, how do you feel about assignment writing at the moment?

- I enjoy it, but I think I spend more time on it than I should. *(year 2, NES)*
- After I finish it, I have a sense of accomplishment. *(Mandarin)*
- I really like writing assignments because you feel a sense of satisfaction when they are done especially if you get a good mark afterwards. *(NES)*
- I like assignment writing, I prefer it to exams. *(NES)*
- I do not enjoy it because I always have to rush to meet the deadline. *(Cantonese)*
- I would have enjoyed if there were not so many to do at one time. *(Cantonese)*
- I've always hated English as I am not strong at it. I still feel the same after doing university assignments but at least they are in a subject I am interested in. *(NES)*

Potential outcomes

- Help UK university lecturers to understand differences between Chinese and British ways of writing.
- Increase knowledge on common chunks in British and Chinese students' writing.
- Assist EAP teachers by identifying features of successful third year student writing in both British and Chinese students' writing.
- Aid materials writers in designing coursebooks.

Issues to consider

- Possible case studies:
  - consider ‘missed chunks’ in students’ writing
  - ask students to identify lexicalised language

- How typical are students who submitted assignments?
  - What about the less proficient ones from either language group?
  - How homogenous is this group of ‘Chinese students’?

- What software?

  Use WordSmith Tools to identify clusters + additional software to identify concgrams & semantic sequences

- What am I looking for?

  Lexicalised language is frequently-used language in different types of writing e.g. year 1 and year 3 of study,

  Different disciplines, Chinese and British students.

  Classify chunks according to function – useful for teaching?

References

- British Academic Written English project. http://www2.warwick.ac.uk/fac/soc/celte/research/bawe
- Hoey, 2005. Lexical Priming. Routledge: New York.
- Leedham, M.E. 2006 'Do I Speak Better?': A longitudinal study of lexical chunking in the spoken language of two Japanese students'. In *The East Asian Learner*. 2(2) http://www.brookes.ac.uk/schools/education/eal/eal-2-2/
- Surveymonkey. https://www.surveymonkey.com
- Scott, M. 2008 WordSmith Tools version 5.0. www.lexically.net//wordsmith/purchasing.htm
- Wray, A. 2002 *Formulaic Language and the Lexicon*, Cambridge University Press: Cambridge.

Acknowledgement

- The data in this study come from the British Academic Written English (BAWE) corpus, which was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800)